2037.0.30.001 - Microdata: Census of Population and Housing, Census Sample File, 2011

Previous ISSUE Released at 11:30 AM (CANBERRA TIME) 12/12/2013

Summary

Contents

CONTENTS

Introduction

INTRODUCTION

ABOUT THE CURF

This product provides information about the release of microdata sample files from the 2011 Census of Population and Housing. Included are details about the methodology, how to use the the sample files, and the conditions of use. Data item lists and information on the quality of the microdata are also provided.

Microdata are the most detailed information available from a Census and are generally the responses to individual questions on the Census Form or data derived from two or more questions. This level of detail is released with the approval of the Australian Statistician.

ABOUT THE MICRODATA

The 2011 Census Sample Files (CSFs) are confidentialised Unit Record Files (CURFs) of Census variables. They contain a small random sample of unidentified occupied private dwellings with their associated family and person records, and a random sample of persons from all non-private dwellings together with a record for the associated non-private dwelling. Subject to the limitations of sample size and the data classifications used, the CURF will enable users to tabulate, manipulate and analyse data to their own specifications.

The 1% Basic CURF contains data on 87,798 dwellings, 93,002 families and 215,597 persons. The 5% Expanded CURF contains data on 422,725 dwellings, 450,038 families and 1,083,585 persons.

The data were collected on Census Night, 9 August 2011.

AVAILABLE PRODUCTS

The following microdata products are available from the Census:

- 1% Basic CURF, available on CD-ROM and through the Remote Access Data Laboratory (RADL) or the ABS Data Laboratory (ABSDL).
- 5% Expanded CURF, available through the Remote Access Data Laboratory (RADL) or the ABS Data Laboratory (ABSDL).

The CURF is not available on CD-ROM to overseas customers.

Unless otherwise indicated the term CURF has been used to apply to both the Basic and Expanded versions of the CSF.

The 5% Expanded CURF contains similar information to the 1% Basic one, except that some items are shown in more detail and some items are included that are not available on the 1% Basic CURF. Both the 1% and 5% CURFs are available in SAS, SPSS and STATA formats.

Further information about these services, and other information to assist users in understanding and accessing microdata in general, is available from the Microdata Entry Page on the ABS website.

Before applying for access, users should read and familiarise themselves with the information contained in this manual and the User Manual: Responsible Use of ABS CURFs (cat. no. 1406.0.55.003).

APPLY FOR ACCESS

To apply for access to the Basic and/or Expanded CURF, register and apply in MiCRO.

FURTHER INFORMATION

Further information about the CSF can be found in this manual:

- A detailed list of data items for the Basic and Expanded CURFs is available under the Downloads tab.
- Information on data quality and definitions can be found on the Explanatory Notes tab.

DATA AVAILABLE ON REQUEST

Data obtained in the Census but not contained on the CURF may be available from the ABS, on request, as statistics in tabulated form.

Subject to confidentiality and sampling variability constraints, special tabulations can be produced incorporating data items, populations and geographic areas selected to meet individual requirements. These are available on request, on a fee for service basis. Contact the National Information and Referral Service on 1300 135 070 or client.services@abs.gov.au for further information.

Methodology

This document was added or updated on 13/12/2013.

METHODOLOGY

SELECTION OF SAMPLE

Data in the Census Sample Files (CSFs) represent 1% and 5% samples of all dwelling, family and person records from the 2011 Census of Population and Housing.

The 1% Basic CSF provides a sample of one private dwelling record in every hundred from the Census, and the associated family and person records. Dwellings with more than six usual residents were removed from the sample to ensure confidentiality of large dwellings (see Large households). For non-private dwellings the sampling is applied to persons present, where one person in every hundred is selected and the associated dwelling records included on the file.

The 5% Expanded CSF provides a sample of one private dwelling in every twenty from the Census, and the associated family and person records. Dwellings with more than eight usual residents were removed from the sample to ensure confidentiality of large dwellings (see Large households). For non-private dwellings the sampling is applied to persons present, where five persons in every hundred are selected and the associated dwelling records included on the file.

The 1% CSF and the 5% CSF also contain corresponding family and person records for the selected private dwellings, as well as a 1 in 100 (or 1 in 20) sample of person records from non-private dwellings. Person, family and dwelling estimates can be obtained from the CSF for private dwellings, but only person level estimates are available for non-private dwellings, and hence for the whole Census population.

Large households
To ensure the confidentiality of large households in occupied private dwellings, the number of persons for each household is restricted to a maximum of six usual residents on the 1% Basic CSF and eight usual residents on the 5% Expanded CSF. Dwellings with more than six usual residents for the 1% Basic CSF, or eight usual residents for the 5% Expanded CSF, have been replaced by dwellings of a similar size and from a similar region that do not

have more than the maximum number of usual residents.

Persons in Other Territories, comprising Jervis Bay, Cocos (Keeling) and Christmas Islands, have been excluded from the sample, as have migratory, shipping and off-shore statistical areas.

Changes in previous CURFs can be found in Changes from previous CSF.

ESTIMATION PROCEDURE

An estimate of the total for an item can be obtained by totalling the item for the CSF and then multiplying the result by 100 for the 1% CSF, or by 20 for the 5% CSF. Note that this estimate of total will not correspond exactly to the total that would be obtained from the full Census, firstly because of the exclusion of large dwellings from the CSF, and secondly because of the sampling error arising due to the CSF containing only a sample of Census records.

Averages from the CSF, such as the proportion of persons falling into a particular category, can be used as an estimate of the corresponding average in the Census. For example, the proportion of Australian-born persons who are students is estimated by the proportion of students observed among Australian-born persons on the CSF. Note that if the denominator of such a proportion is known from the full Census then it can be multiplied by the estimated proportion to give an estimate of the numerator. For example, the total number of Australian-born students could be estimated by multiplying the above proportion by the Australian-born population. This gives an alternative estimate from the CSF (rather than counting the Australian-born students on the 1% CSF and multiplying by 100) that may be preferred in some circumstances, since it is more compatible with the known full-Census count.

RELIABILITY OF ESTIMATES

The sampling error should be taken into account when interpreting estimates from the CSF. A measure of the likely difference between an estimate from the CSF and the corresponding full Census value is given by the standard error (SE) of the estimate. The SE indicates the extent to which an estimate might have varied by chance because only a sample of persons was included. There are about two chances in three that a sample estimate will differ by less than one SE from the full Census value, and about 19 chances in 20 that the difference will be less than two SEs. Another measure of sampling variability is the relative standard error (RSE) which is obtained by expressing the SE as a percentage of the estimate to which it refers.

Non-sampling errors may occur in any enumeration - a full count or a sample - and should not be confused with imprecision due to sampling error, which is measured by the SE. Non-sampling errors in the CSF are differences due to the exclusion of large dwellings, while in the Census as a whole there may be inaccuracies that occur because of imperfections in reporting by respondents, errors made in collection (such as when recording responses) and errors made in processing the Census data. It is not possible to quantify non-sampling error, but every effort is made to reduce it to a minimum. For the following examples, non-sampling error is assumed to be zero. In practice, the potential for non-sampling error adds to the uncertainty in the estimates that is caused by sampling variability.

Standard error calculation

Both CSFs can be treated, for the purposes of standard error calculations, as a simple random sample of dwellings from the private dwelling population. For many purposes the non-private dwelling population has only a minor influence on results, and it is sufficient to include each person counted in a non-private dwelling as a separate 'dwelling' when

calculating standard errors.

Dwelling level estimates

Estimates of the SE of averages for dwelling-level items can be obtained using standard formulae for a simple random sample. These standard error formulae require computing the average value of an item of interest per dwelling on the CSF. The formula for y_{AV}, the estimated average of an item that takes value y_d for dwelling d out of n sampled dwellings in a geographic area, is:

$$y_{AV} = \frac{1}{n}\sum_d y_d$$

where $\sum_d$ represents summing over the n dwellings.

The standard error estimate $SE(y_{AV})$ is given by the following formula:

$$SE(y_{AV}) = \sqrt{\frac{1}{n}\frac{1}{n-1}\sum_d (y_d - y_{AV})^2}$$

The estimate y_{TOT} of the total count for this item, and its corresponding SE estimate $SE(y_{TOT})$, are obtained by multiplying the average per dwelling by the number of dwellings in the geographic area. The number of dwellings is approximated with minimal error by:

$$w \times n$$

where w is the weight (100 on the 1% CSF, and 20 on the 5% CSF)

since the construction of the CSF ensures proportional representation of geographic areas. The formulae are as follows:

$$y_{TOT} = w \times n \times y_{AV}$$

$$SE(y_{TOT}) = w \times n \times SE(y_{AV})$$

Note that the geographic area to be used in these calculations should be the smallest geographic area containing the dwellings in question. For example, estimates for a single state should use state as the geographic area.

Person level estimates

The above formulae can be applied to totals of persons by treating the y_d as person counts within the dwelling i.e. y_d is the number of persons from dwelling d with the characteristic of interest. This makes y_{AV} the average number of persons per dwelling having this characteristic, and y_{TOT} the total number of persons in the geographic area with this characteristic.

Family level estimates

Similarly, estimates for family-level items can be obtained by treating the y_d as family counts within the dwelling i.e. y_d is the number of families from dwelling d with the characteristic of interest, y_{AV} is the average number of families per dwelling having the characteristic, and y_{TOT} is the total number of families in the geographic area with the characteristic.

Example of standard error calculation

The Australian-born population of Australia from the 2011 Census is 15,017,845. As the CSF is based on place of enumeration counts, this figure can be obtained from the Place of Enumeration Profile for Australia on the ABS website. However, as the CSF excludes Other Territories, 1,728 Australian-born persons enumerated in Other Territories on Census Night should be excluded, giving a total of 15,016,117 Australian-born persons. Note that persons enumerated in shipping, migratory or off-shore statistical areas are also excluded from the CSF, but as these counts should be minor, and are not easily accessible from the ABS website, they have been ignored from the calculations.

The 1% CSF estimate of this figure is calculated by taking the 149,324 Australian-born persons on the 1% CSF and multiplying it by 100, giving an estimate of $y_{TOT} = 14{,}932{,}400$. The difference between this figure and the full Census figure of 15,016,117 is due to both the exclusion of large dwellings from the CSF and also the sampling error of the CSF estimate. Note that similar calculations can be carried out using data from the 5% CSF, but using a weight of 20. For simplicity, the remaining examples will be based on the 1% CSF figures.

The simplest way to calculate the SE of this estimate is to produce a file with a single record for each dwelling (treating each person from non-private dwellings as a separate dwelling). On this file, the item y_d should give the number of Australian-born persons in the dwelling. A simple aggregation can be applied to this file to calculate $n = 89{,}866$ (the count of dwelling records), $y_{AV} = 1.6616$ (the mean) and $SE(y_{AV}) = 0.0049$ (the standard error of the mean). These are then used to estimate the total and its SE:

$y_{TOT} = 100 \times n \times y_{AV} = 14{,}932{,}400$ (as calculated previously), and

$SE(y_{TOT}) = 100 \times n \times SE(y_{AV}) = 44{,}143$

This SE calculation suggests that there are about two chances in three that the sample estimate will differ by less than 44,143. from the full-Census value, and about 19 chances in 20 that the difference will be less than 2 x44,143 = 88,286 . The range (14,932,400 - 88,286; 14,932,400 + 88,286) = (14,844,114; 15,020,686) is known as the 95% confidence interval. In this example the Census value of 15,016,117 lies within the confidence interval range.

The estimate is low because the CSF excludes some 1,523 persons from large dwellings, approximately 69% of whom would have been born in Australia. This corresponds to excluding approximately 1,523 x 100 x 69% = 105,087 persons from the estimate. Ignoring the effect that these extra people would have had on the standard error, they would increase the estimate to approximately 15,037,487 Australian-born persons and the 95% confidence interval to (15,037,487 - 88,286; 15,037,487+ 88,286) = (14,949,201; 15,125,773). Whilst the initial confidence interval just covers the true value, the revised confidence interval more

comfortably covers the true population value of 15,016,117.

Users may wish to reproduce these figures using the CSF to ensure that they have interpreted the calculations required correctly.

Clustering of the person sample

For some person-level variables, it may be a reasonable approximation to treat the CSF as a simple random sample of *persons*, even though it is in fact a sample of dwellings. This would involve letting d in the above formulae indicate persons rather than dwellings, and replacing n by the number of persons in the CSF geographic area of interest. Person means and associated standard errors could then be obtained by a standard tabulation package applied to the person-level data.

Unfortunately, doing this will typically give an underestimate of the actual SE. The extent of this underestimation depends on how clustered the variable of interest is within dwellings - that is, on how often similar values of the variable tend to occur together in the same dwelling. The understatement of standard error will be greatest for variables that are highly clustered within dwellings, such as birthplace.

For this reason it would be appropriate, when treating the CSF as a sample of persons, to obtain a measure of the effect of clustering for the variables being investigated. A suitable measure is the design factor (DEFT), given by the ratio of the SE calculated correctly (with dwellings as units) to the SE calculated treating persons as units. Standard errors from the person-level analysis can then be adjusted by this factor.

The SE ignoring clustering will be denoted by $SE_p(y_{TOT})$, with the subscript p indicating that it is calculated at the person level. This can be obtained by taking the person-level CSF and creating a variable taking the value 1 for Australian-born persons and 0 otherwise. Applying a simple tabulation package to this person-level file gives $n_p = 215{,}597$ (the count of person records), $y_{pAV} = 0.6926$ (the mean) and $SE_p(y_{pAV}) = 0.000994$ (the standard error of the mean). These are then used to estimate the total and its SE.

$$y_{TOT} = 100 \times n_p \times y_{pAV} = 14{,}932{,}400$$

(as calculated previously) and

$$SE_p(y_{TOT}) = 100 \times n_p \times SE_p(y_{pAV}) = 21{,}424$$

The design factor is then given as

$$DEFT(y_{TOT}) = \frac{SE(y_{TOT})}{SE_p(y_{TOT})} = 2.06$$

Thus the standard error produced ignoring clustering underestimates the actual standard error by a factor of 2. Users could expect that other totals (eg. for geographic regions) for the variable 'Australian-born' would have a similar design factor.

Standard errors for proportions and differences

Proportions

Simple approximations can be used to estimate the standard error for a ratio of counts. If

y_{TOT_1} and y_{TOT_2} are estimated totals for two nested categories (i.e. category 2 is a subset of category 1) then writing $RSE(y_{TOT}) = \frac{SE(y_{TOT})}{y_{TOT}}$

for the relative standard error gives the following approximation:

$$RSE\left(\frac{y_{TOT_2}}{y_{TOT_1}}\right) = \sqrt{RSE\left(y_{TOT_2}\right)^2 - RSE\left(y_{TOT_1}\right)^2}$$

This formula depends on the two categories being nested, and should not be used for distinct categories.

Differences

If two totals are for distinct categories (eg. in comparing estimates across states) then the difference between two totals has the following SE approximation:

$$SE\left(y_{TOT_2} - y_{TOT_1}\right) = \sqrt{SE\left(y_{TOT_2}\right)^2 + SE\left(y_{TOT_1}\right)^2}$$

While this formula will only be exact for differences between separate and uncorrelated (unrelated) characteristics or sub-populations, it is expected to provide a good approximation for most differences likely to be of interest.

Example of a standard error calculation for a proportion

The number of Australian-born persons who are full-time students (STUP=2) can be estimated by producing a dwelling-level file with a variable y_{2i} giving the number of Australian-born full-time students in the dwelling. Tabulating this variable gives $n = 89,866$, $y_{AV_2} = 0.3797$, $SE(y_{AV_2}) = 0.0027$, so that:

$y_{TOT_2} = 100 \times n \times y_{AV_2} = 3,412,100$,

and

$SE(y_{TOT_2}) = 100 \times n \times SE(y_{AV_2}) = 23,844$

Thus there are an estimated 3,412,100 Australian-born full-time students, with a SE of 23,844 and an

$RSE(Y_{TOT_2}) = \frac{23,844}{3,412,100} \times 100 = 0.70\%$.

The RSE of the estimate of Australian-born persons is

$$RSE(Y_{TOT}) = \frac{44,143}{14,932,400} \times 100 = 0.30\%$$

(see earlier calculations).

Thus the estimated proportion of Australian-born persons who are full-time students is given

by

$$\frac{y_{TOT_2}}{y_{TOT}} = \frac{3,412,100}{14,932,400} = 0.2285$$.

The RSE of this proportion is estimated as

$$RSE\left[\frac{y_{TOT_2}}{y_{TOT}}\right] = \sqrt{RSE(y_{TOT_2})^2 - RSE(y_{TOT})^2} = \sqrt{(0.70)^2 - (0.30)^2} = 0.63\%$$

Thus, the RSE on the estimated proportion of 0.2285 is 0.63% and hence the SE is 0.0014.

Regression estimates

One use of the sample file will be to examine relationships between variables using regression methods. By treating the dwelling as the sample unit, standard regression packages can be used unweighted and the resulting standard errors and test statistics will be good estimates. For example, a regression model could be derived for y_i, the number of persons in the dwelling needing assistance with core activities, against various characteristics $x_{1i}, x_{2i}, \ldots, x_{ki}$ Equation such as x_{1i}, the number of persons in the dwelling aged over 65 years, to fit the linear regression model:

$$y_i = a + b_1 x_{1i} + \ldots + b_k x_{ki}$$

Measures of model fit and of significance of the parameters $a, b_1, \ldots, b_k$ from the standard package will then be appropriate. Unfortunately such a linear model may not adequately describe the relationships between variables at a dwelling level.

If a similar regression is performed treating person as the sample unit, the resulting standard errors and measures of significance could be inaccurate or misleading. This arises because the persons in the sample are clustered within dwellings, and so their responses may be "correlated" or affected by similar influences such as characteristics of the dwelling. The extent to which the measures of significance are affected will depend on how clustered the variable y_i is likely to be within dwellings.

If a person-level analysis is performed, such as a 'logistic analysis' of the probability of a person having a given characteristic, then the effect of clustering should be taken into account when interpreting the outcomes. In particular, SEs are likely to be understated, as discussed in the section Clustering of the person sample, and this will tend to increase the apparent significance of modelled effects.

Techniques are available to perform valid analyses at the person level for a sample that is clustered within dwellings, treating persons as being subject to both person and dwelling effects. These techniques include 'multi-level', 'random effect' and 'mixed' modelling. (Footnote 1 and 2)

By using these techniques, models can be used that do a better job of describing the actual relationships between variables at both person and dwelling level. Statistical packages are widely available to validly perform such analyses.

_____

Footnote 1 Goldstein, H. and Arnold, E, 1995, 'Multilevel Statistical Models', 2nd ed.Halsted Press, New York.
Footnote 2 Snijders Tom A. B. and Bosker Roel J, 1999, 'Multilevel analysis : an introduction to basic and advanced multilevel modelling, SAGE, London.

Using the CSF and File Structures

USING THE CSF AND FILE STRUCTURES

ABOUT THE DATA ITEMS

The full classification structures for all CSF data items can be found in the Census Dictionary, 2011 (cat. no. 2901.0).

Many of the classifications in the CSF have been collapsed and the full listings of the CSF classifications are detailed in the Data items lists in the Downloads tab.

OVERSEAS VISITORS

For the 2011 Census, overseas visitors are separately categorised in standard tabulations (where the table population is 'all persons' and with the exception of the Age, Sex and Marital Status tables). For overseas visitors, the only variables available are Age (AGEP), Sex (SEXP) and Registered Marital Status (MSTP). In all other person variables an Overseas visitor category appears in order to separately designate overseas visitors when compiling tables.

DWELLING INDICATOR FOR PERSONS

The DWIP (Dwelling Indicator for Persons) variable was introduced in 2006 as a way of enabling users of the CSF to more easily distinguish between those people enumerated in private dwellings and those enumerated in non-private dwellings (without the need to link to the household file).

The DWIP variable applies to all persons enumerated in an occupied private dwelling or non-private dwelling. Categories are:

1 Enumerated in an occupied private dwelling
2 Enumerated in a non-private dwelling.

As migratory, off-shore and shipping areas were not included in the sample, there is no `Not applicable' category for this variable.

GEOGRAPHIC AREAS

The CSF contains information on the geographic area of selected dwellings. For 2011, geographic areas in the CSF have been based on the Australian Statistical Geography Standard (ASGS). This replaces the Australian Standard Geographical Classification

(ASGC) used in previous CSFs.

To ensure that the information on the file is not likely to enable identification of a person or household, all areas have been defined using a minimum population size from the full Census data set. For the 1% Basic CSF the minimum population size is 250,000 persons (except for the Northern Territory which has a total population of 234,000 persons). For the 5% Expanded CSF the minimum population size is 124,000 persons. All regions can be aggregated to the state level. Records have been randomly ordered within a region to further reduce the likelihood of individual identification.

Geographic regions have been formed from Statistical Areas Level 4 (1% Basic CURF) and Statistical Areas Level 3 (5% Expanded CURF) and are the basis of the following data items: AREAENUM (Area of enumeration), REGUCP (Region of usual residence on Census night), REGU1P (Region of usual residence 1 year ago) and REGU5P (Region of usual residence 5 years ago) data items. A list of the regions is available in the Downloads tab.

RECORD TYPES AND STRUCTURES

There are three types of records: dwelling, family and person records. For the purposes of the CSF these records are stored in three separate files.

The data in the CSF are hierarchical in structure with one or more families in each dwelling and one or more people in each family. The dwelling, family and person level variables included on the file, and the codes used to describe the values within each variable. A complete list of all data items included on the Basic (1%) and Expanded (5%) CSF are provided in Excel spreadsheets located in the Downloads tab.

The dwelling, family and person records can be linked to each other through their respective record IDs: ABSHID – Dwelling (Household) ID, ABSFID – Family ID, and ABSPID – Person ID.

FILE STRUCTURE

CSF 1% Basic CURF file contents

CSV

These files contain Dwelling, Family and Person Level CURF data in a comma delimited ASCII text format.
CSF11 BD.csv
CSF11BF.csv
CSF11BP.csv

SAS

These files contain Dwelling, Family and Person level data for the CURF in SAS for Windows format:
CSF11BD.sas7bdat contains the Dwelling level data
CSF11BF.sas7bdat contains the Family level data
CSF11BP.sas7bdat contains the Person level data

SPSS

These files contain Dwelling, Family and Person level data for the CURF in SPSS for Windows format:

CSF11BD.sav contains the Dwelling level data
CSF11BF.sav contains the Family level data
CSF11BP.sav contains the Person level data

STATA

These files contain Dwelling, Family and Person level data for the CURF in STATA format:
CSF11BD.dta contains the Dwelling level data
CSF11BF.dta contains the Family level data
CSF11BP.dta contains the Person level data

Information Files

FORMATS.sas7bcat
This file is a SAS library containing formats.

CSF11.SAS
This file contains a SAS program to run the SAS formats.

Important Information CD ROM Census Sample File. PDF
This file contains details to sale and use of ABS microdata.

FREQUENCY FILES
These files contain one-way frequencies of all the data items in an ASCII text format.
CSF11BD_freq.txt
CSF11BF_freq.txt
CSF11BP_freq.txt

CSF 5% Expanded CURF file contents

SAS
These files contain the data for the CURF in SAS for Windows format:
CSF11ED.sas7bdat contains the Dwelling level data
CSF11EF.sas7bdat contains the Family level data
CSF11EP.sas7bdat contains the Person level data

SPSS
These files contain the data for the CURF in SPSS for Windows format:
CSF11ED.sav contains the Dwelling level data
CSF11EF.sav contains the Family level data
CSF11EP.sav contains the Person level data

STATA
These files contain the data for the CURF in STATA format:
CSF11ED.dta contains the Dwelling level data
CSF11EF.dta contains the Family level data
CSF11EP.dta contains the Person level data

Information Files

FORMATS.sas7bcat
This file is a SAS library containing formats.

FREQUENCY FILES
These files contain one-way frequencies of all the data items in an ASCII text format.

CSF11ED_freq.txt
CSF11EF_freq.txt
CSF11EP_freq.txt

CONFIDENTIALISATION OF RECORDS

The CSF is released under the **Census and Statistics Act 1905** which provides that data may be released in the form of unit records where the information is not likely to enable the identification of a particular person. Accordingly there are no names or addresses of respondents on the CSF and other steps have been taken to protect the confidentiality of respondents. These include:

- restricting the data items included on the CSF
- reducing the level of detail shown on the CSF for some data items
- changing some characteristics within individual persons records
- limiting the size of households on the CSF.

Changes from previous CSF

CHANGES FROM PREVIOUS CENSUS SAMPLE FILES

KEY CHANGES

There are three main changes to note in the production of the 2011 Census Sample File (CSF):

- the methodology used for selecting large households
- a new geography
- the methodology used for selecting Ancestry categories.

HOUSEHOLD SIZE

To ensure the confidentiality of large households in occupied private dwellings, the maximum number of persons for each household is restricted to six usual residents on the 1% Basic CSF and eight usual residents on the 5% Expanded CSF. In 2006 this was achieved by removing ‘excess’ person records from large dwellings. After the removal of these records, data items such as Household Composition, Household and Family Income were re-derived to take into account the new structure. For 2011 a simpler approach has been used, where all dwellings with more than six usual residents for the 1% Basic CSF or eight usual residents for the 5% Expanded CSF have been replaced by dwellings of a similar size and from a similar region that do not have more than the maximum number of usual residents.

GEOGRAPHIC AREAS

For 2011, geographic areas in the CSF have been based on the Australian Statistical Geography Standard (ASGS). This replaces the Australian Standard Geographical Classification (ASGC) used in previous CSFs.

For further information on geographic areas refer to Using the CSF and File Structures section.

ANCESTRY

Ancestry data items in the CSF are based on the Australian Standard Classification of Cultural and Ethnic Groups, Second Edition, Revision 1 (which was released in 2011).

In 2006, the categories selected from the Ancestry classification for inclusion in the CSFs were based on the top 20 (1% CSF) and top 30 (5% CSF) ancestry responses from the Census. For 2011, only categories with a minimum of 50,000 (1% CSF) and 30,000 (5% CSF) responses have been reported separately. All other responses have been grouped to broader categories in the classification; 1 digit level for the 1% CSF and 2 digit level for the 5% CSF. This change aims to improve consistency in reported ancestry over time.

CHANGES TO DATA ITEMS

Content of the 2011 CSFs largely duplicates the content of the 2006 CSFs. The main differences are:

- One new data item has been added: NPRD – Number of Persons usually Resident in a Dwelling.
- Three data items have been removed: HIEPPD - Equivalised Household Income for Persons Present (weekly), HIPPD - Household Income for Persons Present (weekly) and FIPPF - Family Income for Persons Present (weekly). These items were derived for the 2006 CSF in response to the method used to select large households. The new methodology used to selected large households for 2011 removed the need to re-derive these variables, and household and family income data will again be reported through the standard Census income data items.
- Four new data items have been added to replace
    - HLRD01 - Housing Loan Repayments (monthly) ranges has been replaced with new Mortgage Repayments (monthly) ranges (MRERD). Some codes have been amended.
    - IND06P - For the 2006 Census, Industry of Employment was coded using the Australian and New Zealand Standard Industrial Classification (ANZSIC) 2006. For the 2011 Census, Industry is classified to the Australian and New Zealand Standard Industrial Classification (ANZSIC), 2006 (Revision 1.0) This mnemonic has been changed for the 2011 Census to INDP.
    - LFS06 - 2006 Labour force status has been changed to LFSP for the 2011 Census.
    - OCC06P - For the 2006 Census, Occupation was coded using the Australian and New Zealand Standard Classification for Occupation (ANZSCO). For the 2011 Census, Occupation is classified to Occupations (ANZSCO), First Edition, Revision 1. This mnemonic has been changed to OCCP for the 2011 Census.
- Categories within data items have been updated. This applies mostly to the ranges for income related variables, but also includes some minor cosmetic changes to labels (for example, where names have changed from singular to plural).

A complete list of all data items included on the Basic (1%) and Expanded (5%) CSF are provided in Excel spreadsheets located in the Downloads tab.

Data item lists

DATA ITEMS LIST

CSF DATA ITEMS

A complete list of all data items included on the Basic (1%) and Expanded (5%) CSF are provided in Excel spreadsheets located in the Downloads tab.

The 1% Basic CSF contains data on

- 87,798 dwellings, 93,002 families and 215,597 persons
- 16 dwelling level, 4 family level and 43 person level data items.

The 5% Expanded CSF contains data on

- 422,725 dwellings, 450,038 families and 1,083,585 persons
- 16 dwelling level, 4 family level and 46 person level data items.

Subject to the limitations of sample size and the data classifications used, the CSF will enable users to tabulate, manipulate and analyse data to their own specifications. Data items are based on the Census Dictionary, 2011 (cat. no. 2901.0).

Conditions of use

CONDITIONS OF USE

USER RESPONSIBILITIES

The Census and Statistics Act includes a legislative guarantee to respondents that their confidentiality will be protected. This is fundamental to the trust the Australian public has in the ABS, and that trust is in turn fundamental to the excellent quality of ABS information. Without that trust, survey respondents may be less forthcoming or truthful in answering our questionnaires. For more information, see 'Avoiding inadvertent disclosure' and 'Microdata' on our web page How the ABS keeps your information confidential .

CURF DATA
The release of CURF data is authorised by clause 7 of the Statistics Determination made under subsection 13(1) of the *Census and Statistics Act 1905*. The release of a CURF must satisfy the ABS legislative obligation to release information in a manner that is not likely to enable the identification of a particular person or organisation.

This legislation allows the Australian Statistician to approve the release of unit record data. All CURFs released have been approved by the Statistician. Prior to being granted access to CURFs, each organisation's Responsible Officer must submit a CURF Undertaking to the

ABS. The CURF Undertaking is required by legislation and states that, prior to CURFs being released to an organisation, a Responsible Officer must undertake to ensure that the organisation will abide by the conditions of use of CURFs. Individual users are bound by the undertaking signed by the Responsible Officer.

All CURF users are required to read and abide by the conditions and restrictions in the User Manual: Responsible Use of ABS CURFs (cat. no. 1406.0.55.003). Any breach of the CURF Undertaking may result in withdrawal of service to individuals and/or organisations. Further information is contained in the Consequences of Failing to Comply web page.

CONDITIONS OF SALE

All ABS products and services are provided subject to the ABS Conditions of Sale. Any queries relating to these Conditions of Sale should be emailed to intermediary.management@abs.gov.au.

PRICE

Microdata access is priced according to ABS Pricing Policy and Commonwealth Cost Recovery Guidelines. For details refer to ABS Pricing Policy on the ABS website. For microdata prices refer to the Microdata prices web page.

HOW TO APPLY FOR ACCESS

Clients wishing to access the microdata should read the How to Apply for Microdata web page. Clients should familiarise themselves with the User Manual: Responsible Use of ABS CURFs (cat. no. 1406.0.55.003) and other related microdata information which are available via the Microdata web pages, before applying for access through MiCRO.

AUSTRALIAN UNIVERSITIES

The ABS/Universities Australia Agreement provides participating universities with access to a range of ABS products and services. This includes access to microdata. For further information, university clients should refer to the ABS Universities Australia Agreement web page.

CITATIONS

Information or data from the Australian Bureau of Statistics must be acknowledged responsibly whenever it is used. Citing, or referencing is important for several reasons, including acknowledging that one has used the ideas, words or data of others. Accurately citing sources used also allows others to find and use the original information. For information on how to cite ABS data please refer to How to cite ABS Sources.

FURTHER INFORMATION

The Microdata Entry Page on the ABS website contains links to microdata related information to assist users to understand and access microdata. For further information users should email microdata.access@abs.gov.au or telephone (02) 6252 7714.

About this Release

The following microdata products are available from the Census Sample File
- Basic CURF on CD-ROM
- Expanded CURF via the Remote Access Data Laboratory (RADL) and ABS Data Laboratory (ABSDL)

Apply online for access to these products at https://www.abs.gov.au/about/microdata.

These products provide sample records for 1% and 5% of the 2011 Census of Population and Housing. A detailed list of data items is available on the Downloads tab.

The microdata enables users to tabulate, manipulate and analyse data. Steps to confidentialise the dataset are taken to ensure the integrity of data and maintain confidentiality of respondents. This includes removing any information that might uniquely identify an individual, reducing the level of detail for some items and collapsing some categories.

Approved users can combine information collected in the Census. The files include 3 levels: Persons, Dwellings, and Families.

Explanatory Notes

Definitions of Quality

DEFINITIONS OF QUALITY

Census Definitions can be found in the Census Dictionary, 2011 (cat. no. 2901.0).

DATA QUALITY

There are a number of resources that explain quality in 2011 Census data. These are the:

- CURF quality declaration
- Census quality declaration
- Census data quality fact sheets, which assist in the use and interpretation of 2011 Census data by providing a summary of conceptual and data issues, and changes that have occurred since the last Census
- Census data quality statements, which are available for all 2011 Census data variables.

Quality Declaration

QUALITY DECLARATION

INSTITUTIONAL ENVIRONMENT

Confidentialised Unit Record Files (CURFs) are released in accordance with the conditions specified in the Statistics Determination section of the Census and Statistics Act 1905 (CSA). This ensures that confidentiality is maintained whilst enabling micro level data to be

released. More information on the confidentiality practices associated with CURFs can be found on the About CURF Microdata page.

For information on the institutional environment of the Australian Bureau of Statistics (ABS), including the legislative obligations of the ABS, financing and governance arrangements, and mechanisms for scrutiny of ABS operations, please see ABS Institutional Environment.

RELEVANCE

Microdata from the 2011 Census of Population and Housing are available as 1% Basic CURF and 5% Expanded CURF. The microdata are the most detailed information available about key characteristics of people in Australia on Census night. These characteristics are generally responses to individual questions on the Census form or data derived from two or more questions.

TIMELINESS

The Census and Statistics Act 1905 requires the Australian Statistician to conduct a Census on a regular basis. Since 1961, a census has been held every five years. The 2011 Census was the 16th national Census, and was held on 9 August 2011. CURFs in recent times have been released within three years of the completion of the Census. Based on the previous schedule the 2016 CURF should be available in 2018/19.

ACCURACY

The microdata generally contains finer levels of detail of data items than what is otherwise published in other formats, for example, in 2011 Community Profiles. For more information on the level of detail provided, see the associated data item listings.

Steps to confidentialise the data made available on the microdata are taken in such a way as to maximise the usefulness of the content while maintaining the confidentiality of respondents selected in the sample. As a result, it may not be possible to exactly reconcile all the statistics produced from the microdata with other published statistics.

COHERENCE

It is important for Census microdata to be comparable and compatible with previous Censuses. There have been minimal changes to the CURFs between 2006 and 2011 to ensure this, however:

- There are differences regarding how the sample has been created in relation to larger households.
- Geographic areas on the 2011 Census Sample Files are based on the Australian Statistical Geography Standard (ASGS), which replaces the Australian Statistical Geography Classification (ASGS) used in previous CURFs.
- Ancestry items (ANC1P and ANC2P) are now based on the Australian Standard Classification of Cultural and Ethnic Groups, Second Edition, Revision 1 (which was released in 2011) and the method of selection of ancestry categories for the 2011 CURF has changed.

The Changes from previous CSF section provides more detailed information on the differences between 2006 and 2011 CURF sample design.

INTERPRETABILITY

The information within this product should be referred to when using the microdata. It contains information including sample methodology, using the CURF and file structure, conditions of use and the data item lists, and changes over time.

The Census Dictionary, 2011 (cat. no. 2901.0) includes information on the Census objectives, methods and design, content, data quality and interpretation, output data items, information about the availability of results and comparability with previous surveys.

ACCESSIBILITY

Microdata products are available to approved users. Users wishing to access the microdata should read the How to Apply for Microdata web page, before applying for access through MiCRO. Users should also familiarise themselves with information available via the Microdata Entry Page.

A full list of available microdata can be viewed via the Expected and available Microdata. More detail regarding types and modes of access to CURFs can be found on the CURF Access Modes and Levels of Detail web page.

The Expanded CURF can be accessed through the Remote Access Data Laboratory (RADL) and the ABS Data Laboratory (ABSDL).

Any questions regarding access to microdata can be forwarded to microdata.access@abs.gov.au or phone (02) 6252 7714.

Glossary

For a list of terms and data items used in the 2011 Census of Population and Housing, refer to Census Dictionary, 2011 (cat. no. 2901.0).

Abbreviations

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ABSDL	ABS Data Laboratory
ANZSCO	Australian and New Zealand Standard Classification of Occupations
ANZSIC	Australian and New Zealand Standard Industrial Classification
ASCCEG	Australian Standard Classification of Cultural and Ethnic Groups
ASCL	Australian Classification of Languages
ASCRG	Australian Standard Classification of Religious Groups
ASGS	Australian Statistical Geography Standard
CSF	Census Sample File
CSV	Comma separated value file
CURF	Confidentialised Unit Record File
HSF	Household Sample File

RADL	Remote Access Data Laboratory
RSE	Relative standard error
SACC	Standard Australian Classification of Countries
SAS	Software package for preparing and executing computerised data analysis
SE	Standard error
SPSS	Software package for preparing and executing computerised data analysis
SA	Statistical Area
STATA	Software package for preparing and executing computerised data analysis